Denominators

QTM 110

ZACHARY BINNEY

OXFORD COLLEGE OF EMORY UNIVERSITY

FALL 2020

Where We Are in the Course

Assorted Quantitative Reasoning Topics

- The Final Chapter

- A bunch of independent but extremely important topics
  - Denominators – TODAY
  - Regression to the Mean
  - Behavioral Adaptation
  - Bayesian Inference
  - Multiple Testing
  - Reporting/Publication Bias
  - Outliers
  - Dangers of *Unthinking/Uncritical* Quantification

Roadmap

- 1. Why Denominators?
  - Counting and Context
- 2. Not Using a Denominator
- 3. Using the Wrong Denominator
- 4. Absolute vs. Relative Difference
- 5. Misleading Visualizations

Why Denominators?

Three Types of Questions

- **1. Descriptive**
  - Describing the world as it *is*

- **2. Predictive**
  - Describing the world as it *will be*

- **3. Causal (Counterfactual Prediction)**
  - Describing the world as it *could be*

Three Types of Questions

- **2. Predictive**

- **3. Causal (Counterfactual Prediction)**

- Denominators are somewhat important here (e.g. ensuring you know what a "1-unit" change in X is, and what a change in Y means), but it's not our focus today…

Three Types of Questions

- **1. Descriptive**
  - Describing the world as it *is*

- **2. Predictive**
  - Describing the world as it *will be*

- **3. Causal (Counterfactual Prediction)**
  - Describing the world as it *could be*

Description and Denominators

- **1. Descriptive**

- The "simplest" type of question? It's both:
  - Extremely common – a bulk of the analysis of the world you'll see and do
  - Much more nuanced than people think
  - Need to spend more time on it than is typically given in analytics courses

Description and Denominators

- Descriptive questions involve "just" <u>counting</u>. Need a:
  - Numerator – number of things
  - Denominator – how many potential things gave rise to the things?

Description and Denominators

- Descriptive questions involve "just" <u>counting</u>. Need a:
  - Numerator – number of things
    - The "easy" part – but sometimes it's hard to just count things
      - COVID-19 cases without testing
      - Police killings without a national database
  - Denominator – how many potential things gave rise to the things?

Description and Denominators

- Descriptive questions involve "just" <u>counting</u>. Need a:
  - Numerator – number of things
  - Denominator – how many potential things gave rise to the things?
    - Make sure 1. all things that *could be* in the numerator are included, 2. but nothing extra
    - Provides **context** – is the numerator big or small?
    - Easy to mess up in good faith…
    - …or manipulate in bad faith to send the wrong message
    - Each question you ask may require a different denominator!

Our focus today for a reason

Description and Denominators

- Descriptive questions involve "just" <u>counting</u>. Need a:
  - Numerator – number of things
  - Denominator – how many potential things gave rise to the things?
    - Today I'm basically going to give you a bunch of examples of some ways denominators get screwed up (or deliberately used for bullshit), so you hopefully will be on the watch for them in the future…
    - Mentally put a **BULLSHIT DETECTION** tag on this whole lecture

Not Using a Denominator

Not Using Denominators

- Sometimes people just ignore denominators

- Numerator: There were 34 cases of COVID-19 last week. Is that bad?
  - Is our denominator New Zealand?
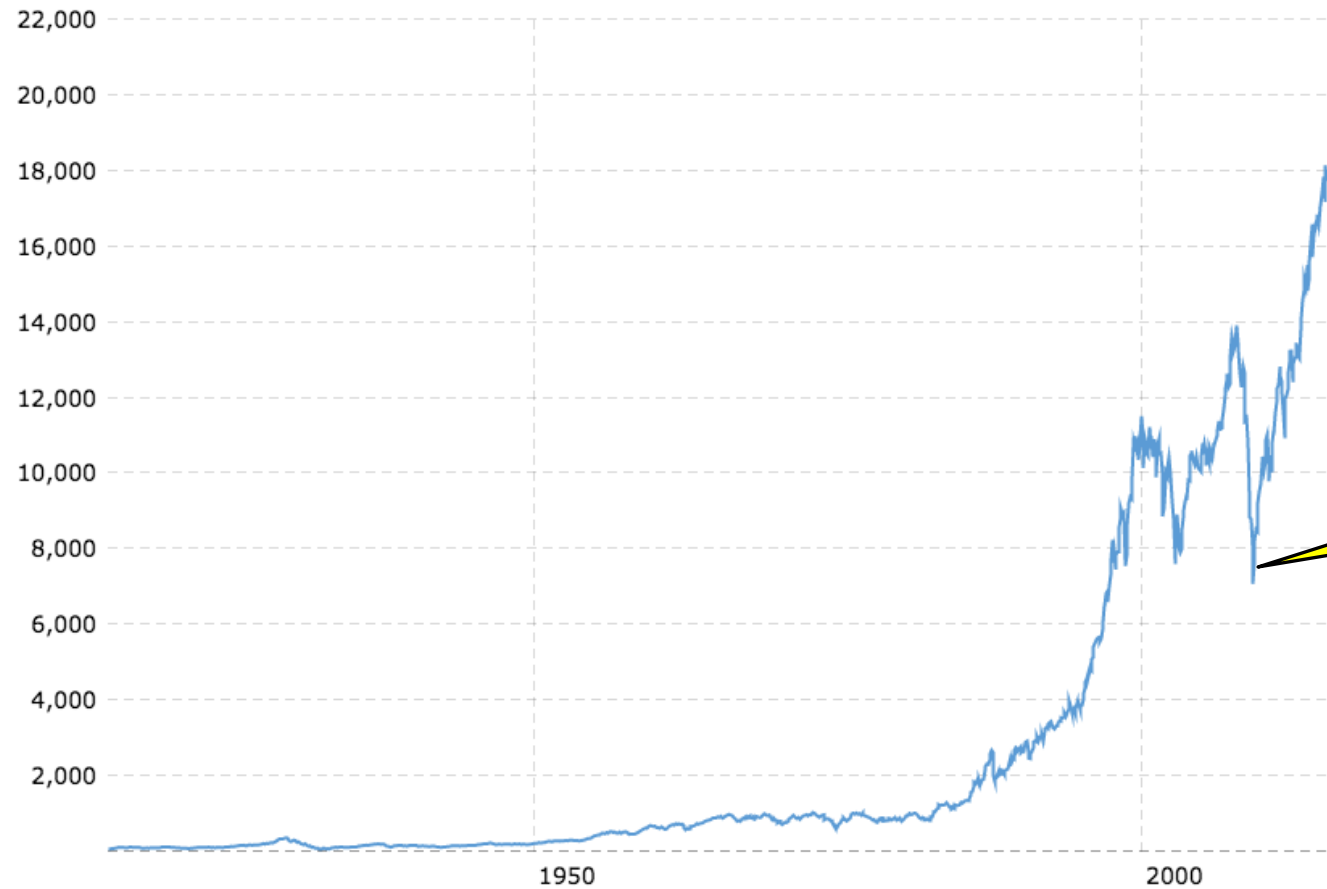 - Or is our denominator the White House?

"Biggest" Stock Market Drops

- Question: What were the "<u>biggest</u>" stock market drops in history?

NOTE: As of 2019 (so, pre-COVID)

Rank	Date	Close	Net change
1	2008-09-29	10,365.45	−777.68
2	2008-10-15	8,577.91	−733.08
3	2001-09-17	8,920.70	−684.81
4	2008-12-01	8,149.09	−679.95
5	2008-10-09	8,579.19	−678.91
6	2011-08-08	10,809.85	−634.76
7	2000-04-14	10,305.78	−617.77
8	2016-06-14	17,400.75	−610.32
9	2015-08-24	15,871.35	−588.40
10	1997-10-27	7,161.14	−554.26

"Biggest" Stock Market Drops

Fun fact: when I graduated college!

"Biggest" Stock Market Drops

- Question: What were the <u>most severe</u> stock market drops in history?

| Rank	Date	Close	Net change	% Change
1	1987-10-19	1,738.74	−508.00	−22.61
2	1929-10-28	260.64	−38.33	−12.82
3	1899-12-18	58.27	−7.94	−11.99
4	1929-10-29	230.07	−30.57	−11.73
5	1929-11-06	232.13	−25.55	−9.92
6	1932-08-12	63.11	−5.79	−8.40
7	1907-03-14	76.23	−6.89	−8.29
8	1987-10-26	1,793.93	−156.83	−8.04
9	2008-10-15	8,577.91	−733.08	−7.87
10	1933-07-21	88.71	−7.55	−7.84

Let's Steal Some Cars

- Well...as Mr. MF asks, what *is* the connection?

Maps Neglecting Denominators

- Please make your maps per capita (i.e. use population denominators
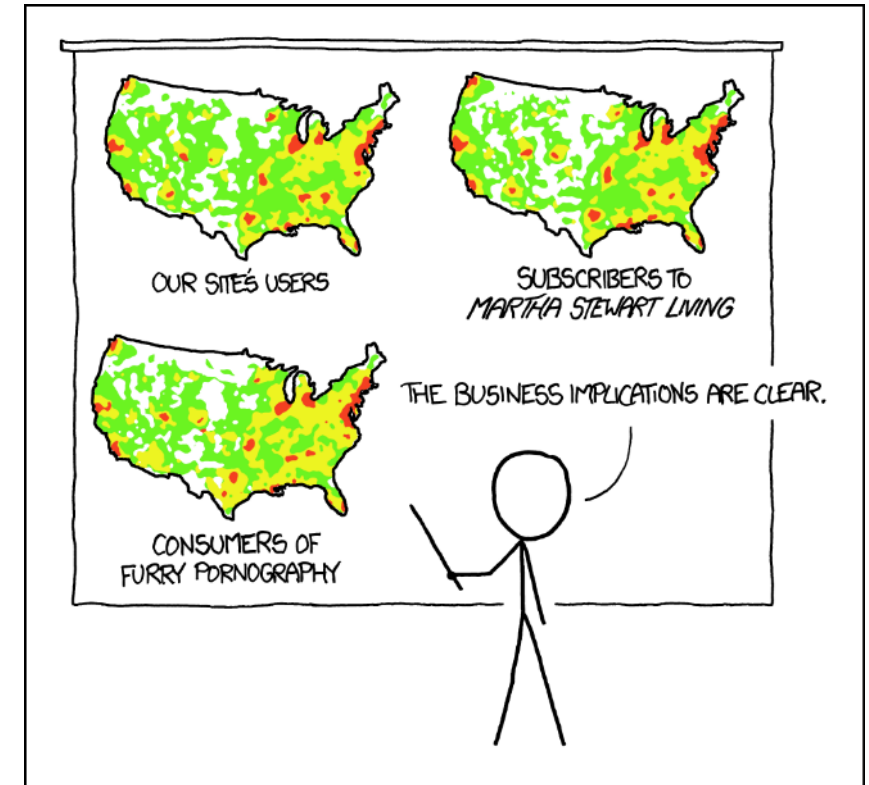
Do You *Always* Want a Denominator?

- Be careful how you use denominators isn't ethically "weird"

  - 1,074 Israelis killed by Palestinians from 2000-05

Country	Proportion	People Killed
Israel	0.015%	1,074
U.S.	0.015%	~50,000
China	0.015%	~223,000
Spain	0.015%	~7,700
Slovenia	0.015%	~300
Tuvalu	0.015%	1-2

- Are these equally "bad" things?

Source: Ellenberg, Jordan. *How Not to be Wrong: The Power of Mathmetical Thinking*; Ch. 4

Do You *Always* Want a Denominator?

- Be careful how you use denominators isn't ethically "weird"

  - 1,074 Israelis killed by Palestinians from 2000-05

Country	Proportion	People Killed
Israel	0.015%	1,074
U.S.	~0.0003%	1,074
China	~0.00007%	1,074
Spain	~0.003%	1,074
Slovenia	~0.05%	1,074
Tuvalu	~10.7%	1,074

  - Are these equally "bad" things?

Do You *Always* Want a Denominator?

- Be careful how you use denominators isn't ethically "weird"

  - As a proportion of population, the 1994 genocide in Rwanda was 9 times as bad as the Holocaust
    - Please do not ever write or say this…
    - They're both horrors that killed 800,000 and 11,000,000, respectively
    - Human lives are human lives – don't discount some because they come from a bigger denominator

  - It's of course useful to know the % of a population affected by something – just use common sense
    - *Sometimes* absolute numbers (numerators) make more sense

Using the Wrong Denominator

Do I Have the Wrong Denominator?

- When selecting a denominator:

1. Ask yourself, "What question is this answering? Is it the one I *should* be answering?"

AND

2. Match your denominator to your numerator!

Do I Have the Wrong Denominator?

- When selecting a denominator:

Ask yourself, "What story is this telling? And is it the one I *should* be telling?"

h/t Mark Simon, Sports Info Solutions

COVID-19 Death Rate in U.S.

- This interview of Donald Trump by Axios in August 2020 – watch the next 60 seconds

COVID-19 Death Rate in U.S.

- [This interview of Donald Trump by Axios in August 2020](#) – watch the next 60 seconds

  - Two different measures of "**death rate**"

  - Trump: $\dfrac{Deaths}{\color{red}{Cases}}$ (Case-fatality rate) – U.S. doing OK!

  - Axios: $\dfrac{Deaths}{\color{yellow}{Population}}$ (COVID-19 Mortality rate) – U.S. doing horribly!

  - **Specific definitions are CRITICAL!!**
  - People can easily bullshit you if they don't specify what they mean by a term – or their denominator. Make them!

COVID-19 Death Rate in U.S.

- [This interview of Donald Trump by Axios in August 2020](#) – watch the next 60 seconds
  - Two different measures of "**death rate**"

  - Trump: $\dfrac{Deaths}{Cases}$ (Case-fatality rate) – U.S. doing OK!

  - Axios: $\dfrac{Deaths}{Population}$ (COVID-19 Mortality rate) – U.S. doing horribly!

  - Which tells the more important story?

COVID-19 Case Rate in NFL

- In NFL in 2020, all players & staff tested daily

- Sports leagues and media struggle with reporting results. Two common options:
  - 1. $\dfrac{99 \; positives}{\sim 400{,}000 \; tests} = 2.5 \; per \; 10{,}000$
  - 2. $\dfrac{99 \; positives}{7{,}820 \; people} = 126.6 \; per \; 10{,}000$

  - Which is better?
    - What's the question we want to answer/story we want to tell?
    - Maybe something like "What is the burden of COVID-19 in the NFL (vs. the general population)?"

Kevin Seifert @SeifertESPN

Here are the NFL's COVID-19 testing results from the most recent period (Oct. 4-10):

Monitoring Testing results for October 4 – October 10:
- 37,912 tests were administered to a total of 7,820 players and team personnel.
- 14,713 tests were administered to 2,456 players; 23,199 tests were administered to 5,364 personnel.
- There were **eight** new confirmed positive tests among players and **seven** new confirmed positives among other personnel.

Total test results for the entire Monitoring Testing period to-date, Aug 1 – Oct 10:
- During Monitoring Testing from Aug 1 – Oct 10, **39** players and **60** other personnel were confirmed positive cases.
- More than **400,000** tests were administered to players and personnel during the Aug 1 – Oct 10 period.

1:02 PM · Oct 13, 2020 · Twitter Web App

COVID-19 Case Rate in NFL

- In NFL in 2020, all players & staff tested daily

- Sports leagues and media love struggle with reporting results. Two common options:

  - 1. $\frac{99\ positives}{\sim 400,000\ tests} = 2.5\ per\ 10,000$

  - 2. $\frac{99\ positives}{7,820\ people} = 126.6\ per\ 10,000$

- What if I tested everyone 3x/day instead? Has the situation changed?

  - 1. $\frac{99\ positives}{\sim 1,200,000\ tests} = 0.8\ per\ 10,000$

  - 2. $\frac{99\ positives}{7,820\ people} = 126.6\ per\ 10,000$

- First denominator answers the wrong question!

  - Basically says "we're testing a lot" – great, but not our question here!

Kevin Seifert ✔
@SeifertESPN

Here are the NFL's COVID-19 testing results from the most recent period (Oct. 4-10):

Monitoring Testing results for October 4 – October 10:

- 37,912 tests were administered to a total of 7,820 players and team personnel.

- 14,713 tests were administered to 2,456 players; 23,199 tests were administered to 5,364 personnel.

- There were **eight** new confirmed positive tests among players and **seven** new confirmed positives among other personnel.
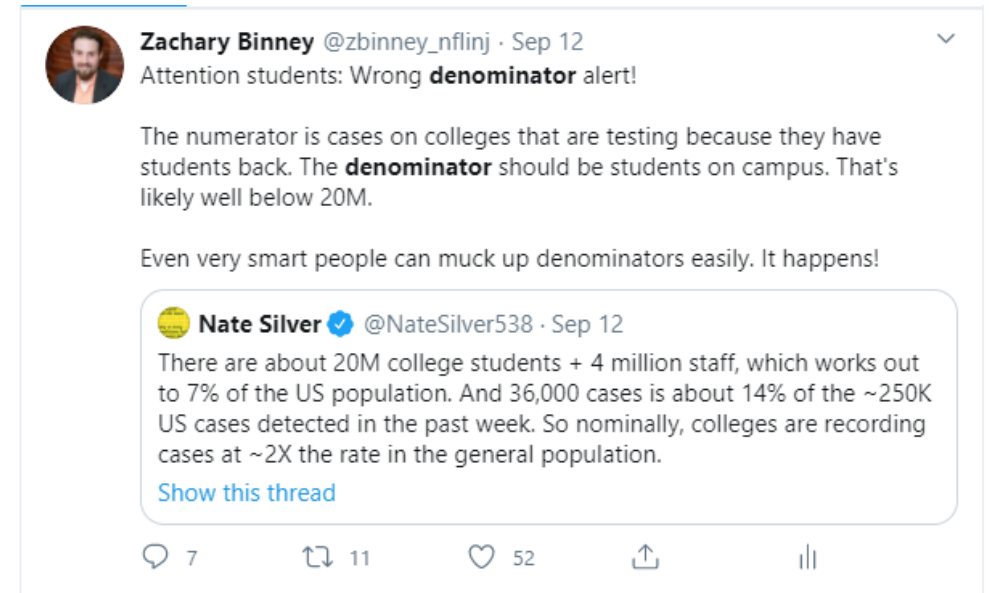
Total test results for the entire Monitoring Testing period to-date, Aug 1 – Oct 10:

- During Monitoring Testing from Aug 1 – Oct 10, **39** players and **60** other personnel were confirmed positive cases.

- More than **400,000** tests were administered to players and personnel during the Aug 1 – Oct 10 period.

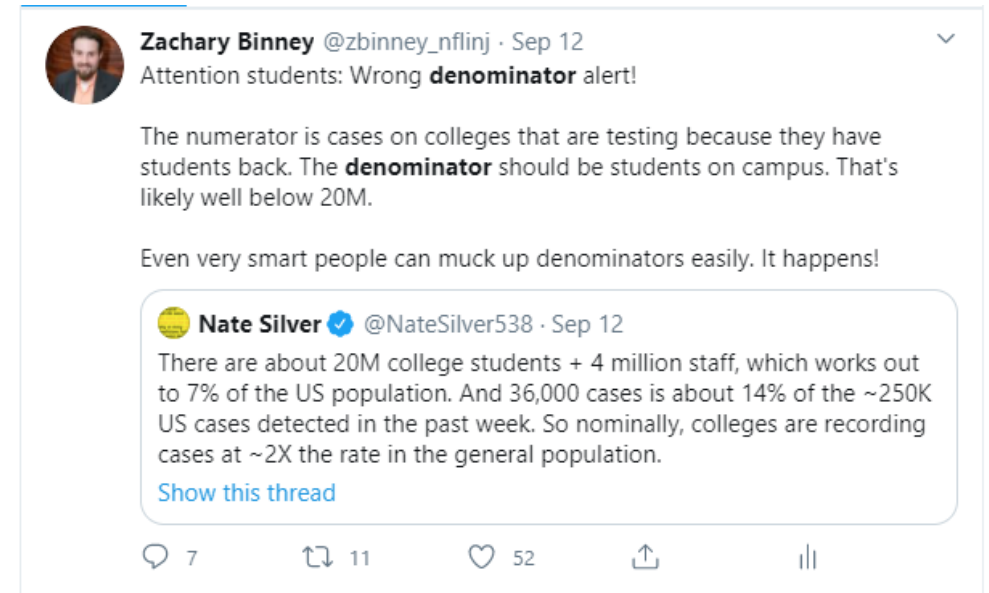1:02 PM · Oct 13, 2020 · Twitter Web App

Match Your Denominator to Your Numerator

- Remember our definitions?
  - Numerator – number of things
  - Denominator – how many potential things gave rise to the things?
    - Make sure 1. all things that could be in the numerator are included, 2. but nothing extra

- Really easy to screw this up!

- Nate does the following:
  - Numerator: COVID-19 cases in colleges with students on-campus who are testing
  - Denominator: <u>all</u> college students + staff

- Can you articulate why this is unfair, and in what direction? What would a better denominator be?

Zachary Binney @zbinney_nflinj · Sep 12
Attention students: Wrong **denominator** alert!

The numerator is cases on colleges that are testing because they have students back. The **denominator** should be students on campus. That's likely well below 20M.

Even very smart people can muck up denominators easily. It happens!

Nate Silver ✓ @NateSilver538 · Sep 12
There are about 20M college students + 4 million staff, which works out to 7% of the US population. And 36,000 cases is about 14% of the ~250K US cases detected in the past week. So nominally, colleges are recording cases at ~2X the rate in the general population.

Show this thread

💬 7   ↻ 11   ♡ 52

Match Your Denominator to Your Numerator

- Nate does the following:
  - Numerator: COVID-19 cases in colleges with students on-campus who are testing
  - Denominator: <u>all</u> college students + staff

- Can you articulate why this is unfair, and in what direction? What would a better denominator be?
  - Denominator is too big – <u>includes people who couldn't be in the numerator</u> (because they're not on campus). That's bad!

  - Causes the rate on college campuses to be *underestimated*

  - Better denominator: number of people on campuses that are 1. open, and 2. testing
    - **Match your denominator to your numerator!**

Zachary Binney @zbinney_nflinj · Sep 12
Attention students: Wrong **denominator** alert!

The numerator is cases on colleges that are testing because they have students back. The **denominator** should be students on campus. That's likely well below 20M.

Even very smart people can muck up denominators easily. It happens!

Nate Silver @NateSilver538 · Sep 12
There are about 20M college students + 4 million staff, which works out to 7% of the US population. And 36,000 cases is about 14% of the ~250K US cases detected in the past week. So nominally, colleges are recording cases at ~2X the rate in the general population.

Show this thread

💬 7   🔁 11   ♡ 52

Observation Selection Effects

- Colleges often put their "average class size" in brochures – but what does that mean?

- Two possible questions:

  - 1. What is the average class size at the college?
    - $\dfrac{Total\ class-enrollments}{Total\ Classes}$ 1 class-enrollment = 1 student in 1 class

  - 2. What will a random student's average class size be, on average?
    - $\dfrac{\sum_{all\ class\ sizes} Total\ class-enrollments * class\ size}{Total\ class-enrollments} =$ Weighted average class size

Observation Selection Effects

- Colleges often put their "average class size" in brochures – but what does that mean?

- Consider a Biology Department:
  - 20 classes with 20 students, serving 400 total students
  - 4 classes with 200 students, serving 800 total students

- Average class size: $\frac{(20*20)+(4*200)}{24} = 50$

- Average *experienced* class size: $\frac{(800*200)+(400*20)}{1200} = 140$

- Key insight: more students will be in large classes than small classes because large classes...have more students!

Observation Selection Effects

- This is a much broader problem than college brochures

  - Like Berkson's bias, you'll start noticing it everywhere…

  - It explains why you feel like you always wait longer for buses (more likely to arrive during long delay than short gap), or why you're always in the slow lane on the highway (more cars in the slower lane…that's why it's slower!), and why all your friends tend to have more friends than you (you're more likely to be friends with someone who has a lot of friends!)

- Note this is different from "selection on observables" in econometrics – that's "no unmeasured confounding"

Careful When Numbers Can be Negative

- Proportions can be misleading when numbers can be negative

U.S. Jobs Gained or Lost by Sector, 2000-2008
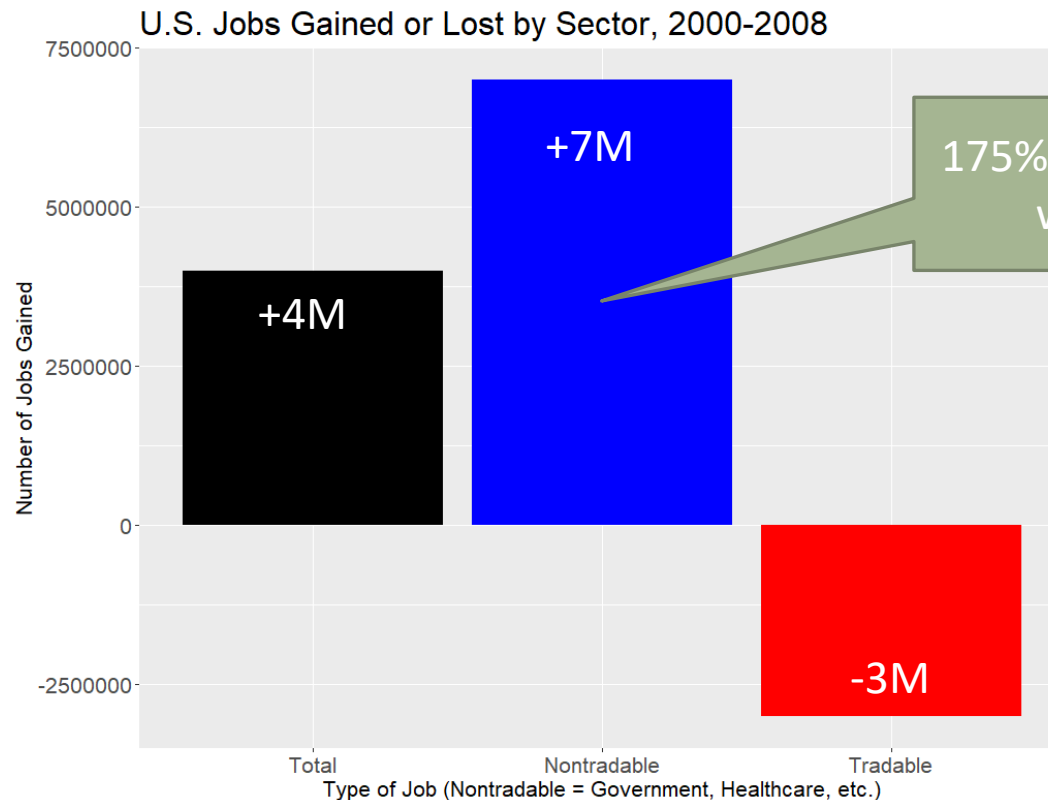
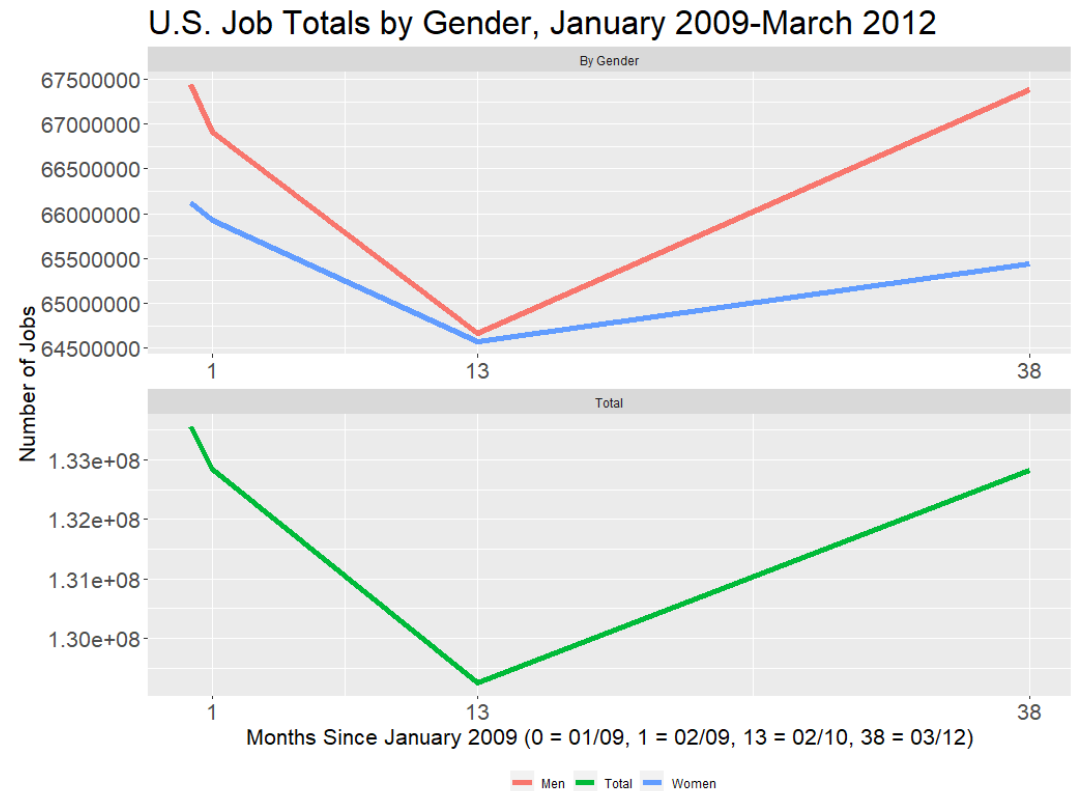175% of the jobs gained from 2000-2008 were in the nontradable sector

Careful When Numbers Can be Negative

- Proportions can be misleading when numbers can be negative
  - Romney 2012 campaign: women account for 92.3% of all jobs lost under Obama
  - Numerator:
    - From January 2009-March 2012, women lost 683,000 jobs
  - Denominator:
    - From January 2009-March 2012, U.S. lost 740,000 jobs
  - Women represented 683,000/740,000 = 92% of job losses
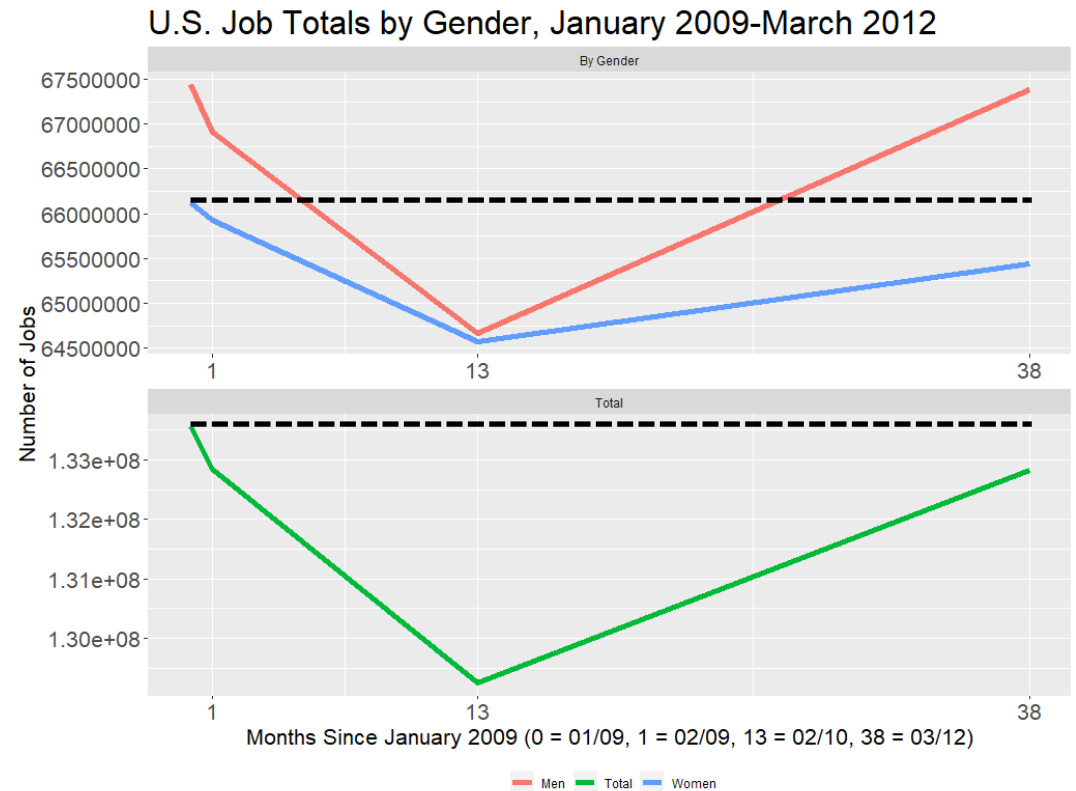
U.S. Job Totals by Gender, January 2009-March 2012

Months Since January 2009 (0 = 01/09, 1 = 02/09, 13 = 02/10, 38 = 03/12)

Source: Ellenberg, Jordan. *How Not to be Wrong: The Power of Mathmetical Thinking*; Ch. 5

Careful When Numbers Can be Negative

- Proportions can be misleading when numbers can be negative
  - Romney 2012 campaign: women account for 92.3% of all jobs lost under Obama
  - Numerator:
    - From February 2009-March 2012, women lost 484,000 jobs
  - Denominator:
    - From February 2009-March 2012, U.S. lost 16,000 jobs
  - Women represented 484,000/16,000 = >3,000% of total job losses!
    - Obviously bullshit

U.S. Job Totals by Gender, January 2009-March 2012

Source: Ellenberg, Jordan. *How Not to be Wrong: The Power of Mathmetical Thinking*; Ch. 5
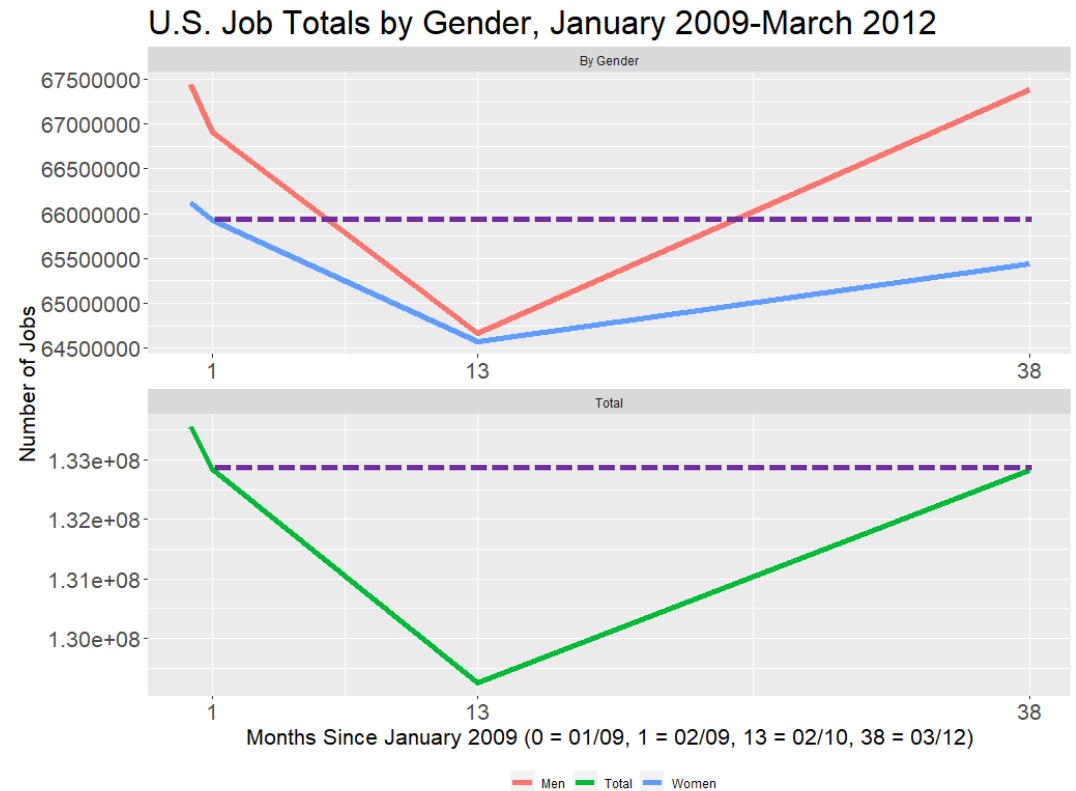
Careful When Numbers Can be Negative

- No question women experienced a weaker recovery than men

- Also no question the Romney campaign's claim is *mathematically 100% correct*

- Right answer to the **WRONG QUESTION**!
  - "True, but false"
  - True mathematically, but implies something untrue
  - Formulating the correct question is much harder than any computations you will do
    - The latter a calculator can do!

- What did you really want to know about the recession?

U.S. Job Totals by Gender, January 2009-March 2012

Months Since January 2009 (0 = 01/09, 1 = 02/09, 13 = 02/10, 38 = 03/12)

Men — Total — Women

Absolute vs. Relative Differences

Absolute Versus Relative Numbers

- Absolute Numbers:
  - 20
  - 500

- Relative Numbers:
  - Proportion: how much of a whole is some number?
    - 500 cases in 1000 people $\rightarrow \frac{500}{1000} = 50\%$
  - Ratio: any comparison of two numbers
    - $\frac{1000}{500} = 2 \; or \; 2:1$ ("2 to 1")

Absolute Versus Relative *Differences*

- Say your exposed group has some outcome (death from heart disease) 20% of the time, and the unexposed group 10% of the time

- Absolute Difference:
  - $\frac{20}{100} - \frac{10}{100} = 10\%$

- Relative Difference:
  - 20%/10% = 2.0

- Say your exposed group has some outcome (death from lung cancer) 2% of the time, and the unexposed group 1% of the time

- Absolute Difference:
  - $\frac{2}{100} - \frac{1}{100} = 1\%$

- Relative Difference:
  - 2%/1% = 2.0

Relative differences can be used to make tiny things seem big…

…or absolute differences can be used to make big things seem tiny!

Need to decide which makes more sense on a case-by-case basis

Why Is This in the Denominators Lecture?

- Say your exposed group has some outcome (death from heart disease) 20% of the time, and the unexposed group 10% of the time

- Absolute Difference:
  - $\frac{20}{100} - \frac{10}{100} = 10\%$

  > Denominator(s) = Total Units in Each Group, or Units That Might be Treated. Don't have to be the same! Ex.: 20/100 – 1000/10,000 = 10%

- Relative Difference:
  - 20%/10% = 2.0

  > Denominator = risk in unexposed/untreated group

Birth Control and Blood Clot Study Results

- [Here's a drag queen dissecting misleading headlines on TikTok](), if that's your thing

- Headlines: "Birth Control X doubles your risk of blood clots!"

- Actual study? →

- Reporting has real impacts – women dropped this birth control, which may have led to a rise in unplanned pregnancies

onlinekyne ● Kyne · 5-13
Numbers don't have to be wrong to be misleading. #math #education #edutok #drag #dragqueen
♫ original sound - Kyne

15 in 100,000 = 0.015%
25 in 100,000 = 0.025%

♡ 43.4K   💬 343   ↪ 452

onlinekyne ● Kyne · 5-13
Numbers don't have to be wrong to be misleading. #math #education #edutok #drag #dragqueen
♫ original sound - Kyne

Risk of blood clots almost twice as likely

Risk of blood clots rises from 0.015% to 0.025%

♡ 43.4K   💬 343   ↪ 452

Homer Simpson and Crime

- *The Simpsons*, "Homer the Vigilante"
  - After a burglary, Homer starts a "posse" to deal with crime in Springfield
  - Is overall crime up or down?

Evan @ereller25 · Sep 27

Mr. Simpson, how do you respond to the charge that petty vandalism such as graffiti is down 80%, while heavy sack beatings are up a shocking 900%?

Kent Brockman : Mr. Simpson, how do you respond to the charges that petty vandalism such as graffiti is down eighty percent, while heavy sack beatings are up a shocking nine hundred percent?

Homer Simpson : Aw, you can come up with statistics to prove anything, Kent. Forfty percent of all people know that.

Kent Brockman : I see. Well, what do you say to the accusation that your group has been causing more crimes than it's been preventing?

Homer Simpson : Oh, Kent, I'd be lying if I said my men weren't committing crimes.

Kent Brockman : *[pause]* Well, touché.

Be Careful Around 0% and 100%

- One place absolute vs. relative differences become important is near the extremes of probabilities (0% or 100%)
  - Express confidence two different but equivalent ways:
    - 97% vs. 99% sure X will happen – relative difference tiny
    - 3% vs. 1% sure X *won't* happen – relative difference enormous
  - Which *feels* like a bigger difference? You have to be on guard for this!

Nate Silver ✓ @NateSilver538 · Aug 14

Some stuff I agree with here, but I *don't* agree with the idea that an 89% forecast (Economist) and a 72% forecast (538) are close.

Maybe easier to see if you compare a 28% chance of Trump winning to 11%; we think a Trump win is 2-3x as likely.

statmodeling.stat.columbia.edu/2020/08/14/new...

💬 82   � 77   ♡ 640
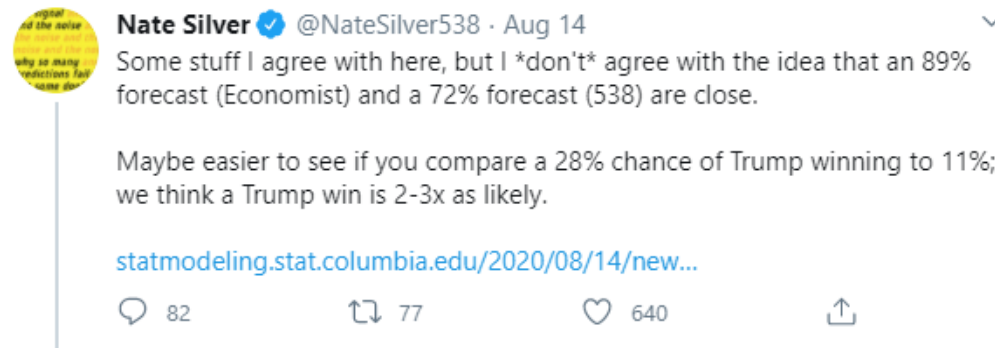
Absolute Versus Relative Differences

- One common manifestation of this problem that merits special discussion is…

Percentage Point vs. Percent Differences

- Say your exposed group has some outcome (death from heart disease) 20% of the time, and the unexposed group 10% of the time

- **Percentage Point Difference**:
  - 20% - 10% = exposed group is 10 <u>percentage points</u> more likely to die from heart disease

- **Percent Difference**:
  - 20%/10% = exposed group is 100% more likely to die from heart disease

- Say your exposed group has some outcome (death from lung cancer) 2% of the time, and the unexposed group 1% of the time

- Percentage Point Difference:
  - 2% - 1% = exposed group is 1 <u>percentage point</u> more likely to die from lung cancer

- Percent Difference:
  - 2%/1% = exposed group is 100% more likely to die from lung cancer

Sometimes will see "200% more likely" – more correct would be "have 200% the risk of the unexposed". Be careful!

GDP Reporting

- Country A's GDP rose 0.1%, but country B's rose 1.0%

- Consider what happens if A's drops just a bit in absolute terms to 0.001%
  - Now Country B is 10,000 times Country A!
  - Always illustrative to try making tiny tweaks to your numbers and see if your interpretations change dramatically – if they do, re-think your presentation

Growth in Country B is 0.9 percentage points higher than Country A

Growth in Country B is 1,000% that of Country A!!!

Misleading Visualizations

Absolute Versus Relative Numbers

- Three denominator-related issues for visualizations:

  - 1. Make sure visualizations include denominators
    - Population maps example

  - 2. Don't plot shifting changes over time

  - **3. Always check axes to see how big changes are!**

Don't Plot Shifting Changes Over Time

- Plotting changes from point *t-1* to *t* over time can be incredibly misleading
  - Really an issue of using the wrong denominator!
  - Advice: plot either 1. Absolute numbers <u>or</u> 2. Cumulative change from a single referent point (e.g. August 2018)
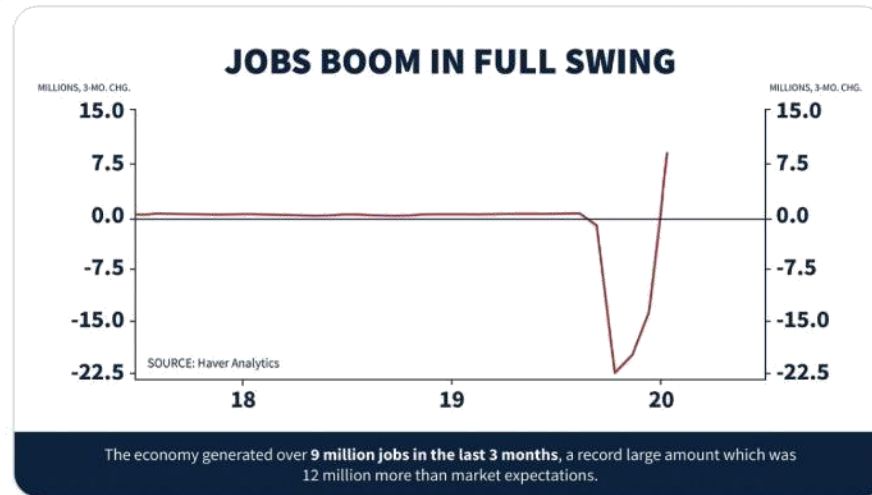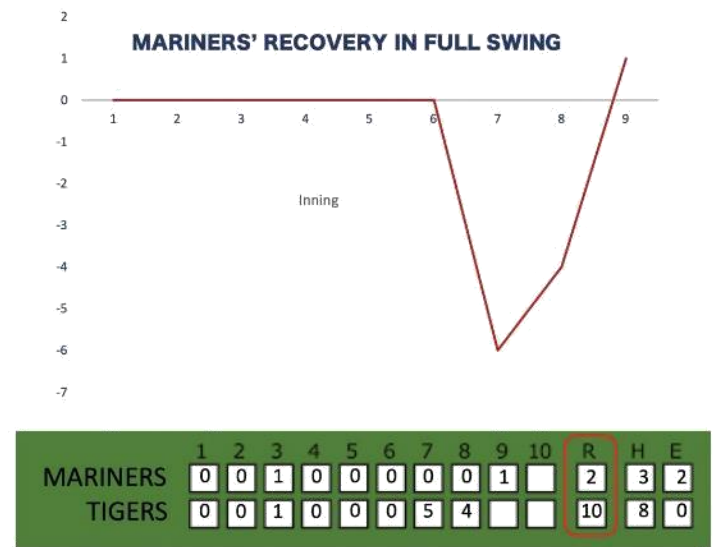
Check Your Axes: Dow Jones Industrial Average

- Dow has a.) changed with inflation and b.) increased the number of companies included (denominator) over time!
  - Started with 12, now 30
  - GE only original left
  - Dropped examples: US Leather, Distilling & Cattle Feeding; American Cotton Oil; American Sugar
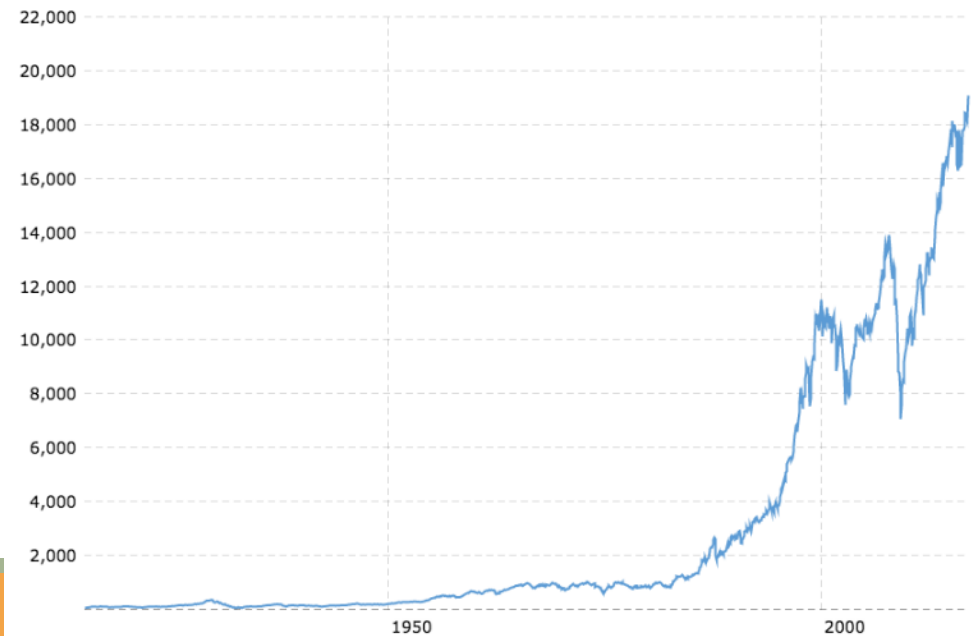
Relative vs. Absolute Changes in Sports

- Are the differences in each of these graphs meaningful?
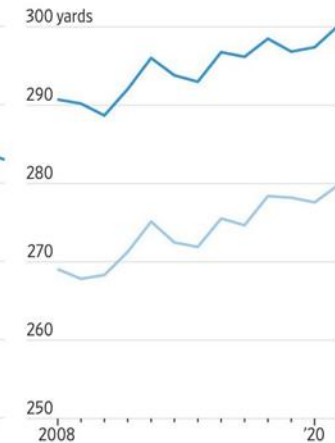
Heights by Country

- Maybe a relative measure would've been a better choice here?
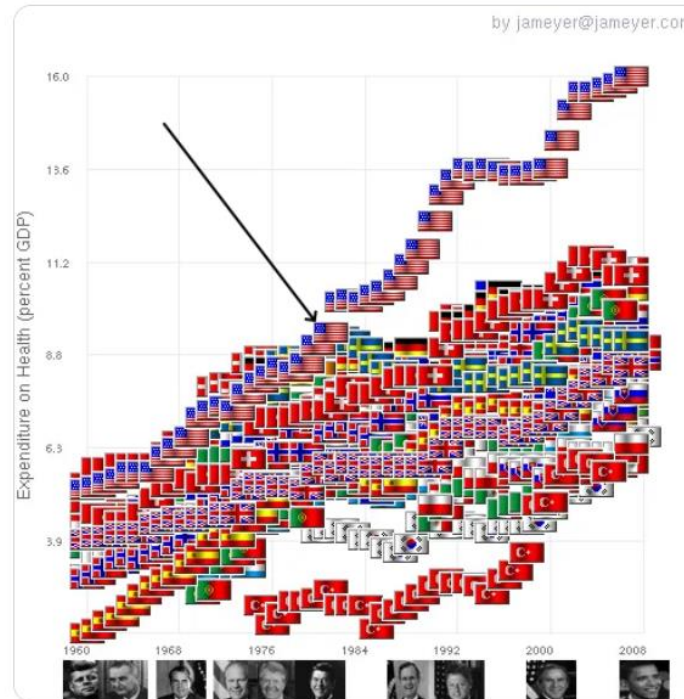
- (Actually, should've just had Y-axis start at zero)

Average Female Height per country
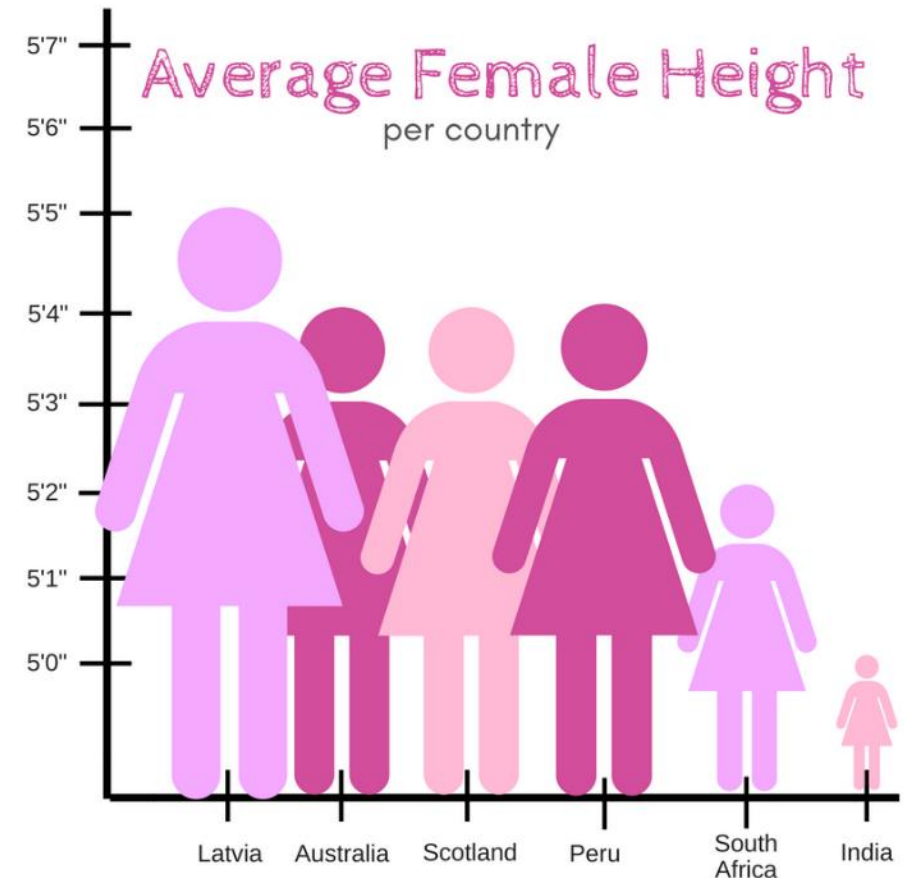
Visualizations Can be Bad for Reasons Other Than Denominators, Too

- This one is just for fun

Visualizations Can be Bad for Reasons Other Than Denominators, Too

- Bad/misleading data visualization is a *huge* problem

- For a more general discussion of bad/misleading data visualization practices I HIGHLY recommend Chapter 7 of *Calling Bullshit* by Carl Bergstrom and Jevin West (only $22!)

- Or their visualization discussions here

- For just a few examples of *good* visualizations, here

Summary

- Using a denominator is very important

- Using the right denominator is even more important

- Does it answer the right question? Does it tell the story it should?

- Match your denominator to your numerator!

- Did the analyst explicitly define their denominator/what terms they're using, if there isn't a single agreed-upon definition?

- Is the analyst – which might be you! – using relative rather than absolute measures to make tiny differences look big – or vice-versa?

- Does a plot use misleading axes?

- **SO MUCH BULLSHIT arises from mis-used denominators!**

Thanks!

- Questions? zbinney@emory.edu, @binney_z on Twitter